Descriptive Statistics: Introduction

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms and boxplots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Introduction

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics"**. You will learn to calculate, and even more importantly, to interpret these measurements and graphs.

Descriptive Statistics: Displaying Data

This module provides a brief introduction into the ways graphs and charts can be used to provide visual representations of data.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample. The graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly.

Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar chart, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), pie charts, and the boxplot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs and bar graphs. Our emphasis will be on histograms and boxplots.

Descriptive Statistics: Stem and Leaf Graphs (Stemplots)
This module introduces the use of stem-and-leaf graphs (stemplots), line graphs and bar graphs for describing a set of data visually.

One simple graph, the **stem-and-leaf graph** or **stem plot**, comes from the field of exploratory data analysis.It is a good choice for numerical data sets that are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem 2 and leaf 3. Four hundred thirty-two (432) has stem 43 and leaf 2. Five thousand four hundred thirty-two (5,432) has stem 543 and leaf 2. The decimal 9.3 has stem 9 and leaf 3. Write the stems in a vertical line from smallest the largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example:
For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
334249495355556163676868696972737478808388888890929494949496
100

Stem	Leaf
3	3
4	299
5	355
6	1378899

Stem	Leaf
7	2348
8	03888
9	0244446
10	0

Stem-and-Leaf Diagram

The stem plot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% of the scores were in the 90's or 100, a fairly high number of As.

The stem plot is a quick way to graph and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers. In the example above, there were no outliers.

Example:
Create a stem plot using the data:
1.11.52.32.52.73.23.33.33.53.84.0 4.24.54.54.74.85.55.66.56.712.3
The data are the distance (in kilometers) from a home to the nearest supermarket.

Exercise:

Problem:

1. Are there any values that might possibly be outliers?
2. Do the data seem to have any concentration of values?

Note: The leaves are to the right of the decimal.

Solution:

The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 kilometers.

Stem	**Leaf**
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	

Stem	**Leaf**
9	
10	
11	
12	3

Glossary

Outlier
: An observation that does not fit the rest of the data.

Descriptive Statistics: Line Graphs and Bar Graphs

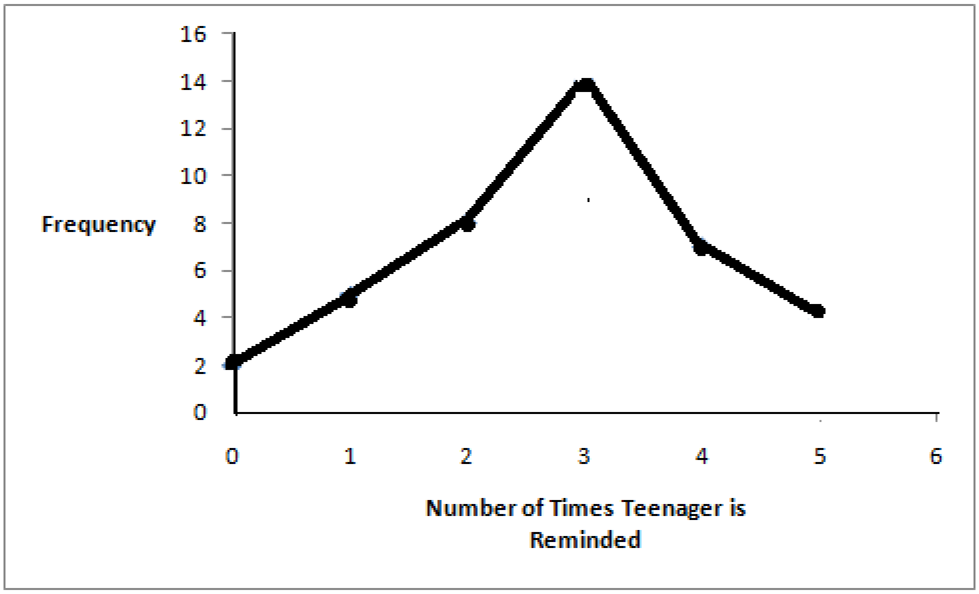
This module introduces the use of stem-and-leaf graphs (stemplots), line graphs and bar graphs for describing a set of data visually.

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in the example, the **x-axis** consists of **data values** and the **y-axis** consists of **frequency points**. The frequency points are connected.

Example:
Line Graphs

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his/her chores. The results are shown in the table and the line graph.

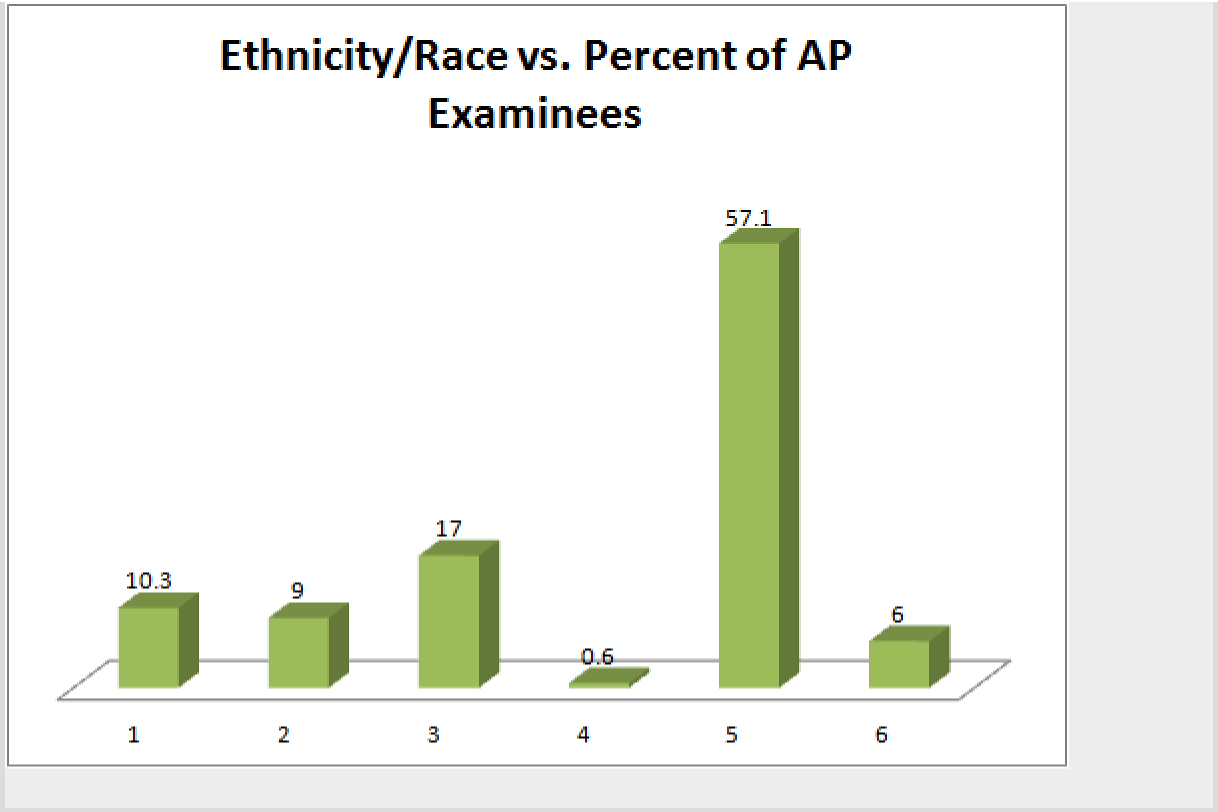
Number of times teenager is reminded	**Frequency**
0	2
1	5
2	8
3	14
4	7
5	4



Bar graphs are useful for displaying categorical data. Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes and they can be vertical or horizontal.

Example:
Bar Graphs
The columns in the table below contain the race/ethnicity of U.S. Public Schools: High School Class of 2011, percentages for the Advanced Placement Examinee Population for that class and percentages for the Overall Student Population. The 3-dimensional graph shows the Race/Ethnicity of U.S. Public Schools (qualitative data) on the **x-axis** and Advanced Placement Examinee Population percentages on the **y-axis**. (**Source: http://www.collegeboard.com** and **Source: http://apreport.collegeboard.org/goals-and-findings/promoting-equity**)

Race/Ethnicity	**AP Examinee Population**	**Overall Student Population**
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%



Go to Outcomes of Education Figure 22 for an example of a bar graph that shows unemployment rates of persons 25 years and older for 2009.

Note: This book contains instructions for constructing a **histogram** and a **box plot** for the TI-83+ and TI-84 calculators. You can find additional instructions for using these calculators on the Texas Instruments (TI) website.

Glossary

Outlier

An observation that does not fit the rest of the data.

Descriptive Statistics: Histogram
This module provides an overview of Descriptive Statistics: Histogram as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous boxes (the boxes touch, unlike in a bar graph). It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **Frequency** or **relative frequency**. The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data. (The next section tells you how to calculate the center and the spread.)

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (In the chapter on Sampling and Data, we defined frequency as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:
Equation:

$$\text{RF} = \frac{f}{n}$$

For example, if 3 students in Mr. Ahab's English class of 40 students received from 90% to 100%, then,

$f = 3$, $n = 40$, and RF $= \frac{f}{n} = \frac{3}{40} = 0.075$

Seven and a half percent of the students received 90% to 100%. Ninety percent to 100 % are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of from 5 to 15 bars or classes for clarity. Choose a starting point for the first interval to be less than the smallest data value.

Example:
The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.
60 60.5 61 61 61.5
63.5 63.5 63.5
64 64 64 64 64 64 64 64.5 64.5 64.5 64.5 64.5 64.5 64.5 64.5
66 66 66 66 66 66 66 66 66 66 66.5 66.5 66.5 66.5 66.5 66.5 66.5 66.5
66.5 66.5 66.5 67 67 67 67 67 67 67 67 67 67 67 67 67.5 67.5 67.5 67.5
67.5 67.5 67.5
68 68 69 69 69 69 69 69 69 69 69 69 69.5 69.5 69.5 69.5 69.5
70 70 70 70 70 70 70.5 70.5 70.5 71 71 71
72 72 72 72.5 72.5 73 73.5
74
The smallest data value is 60. We will use that as our starting point.

Note: We will make each bar or class interval 2 units wide.

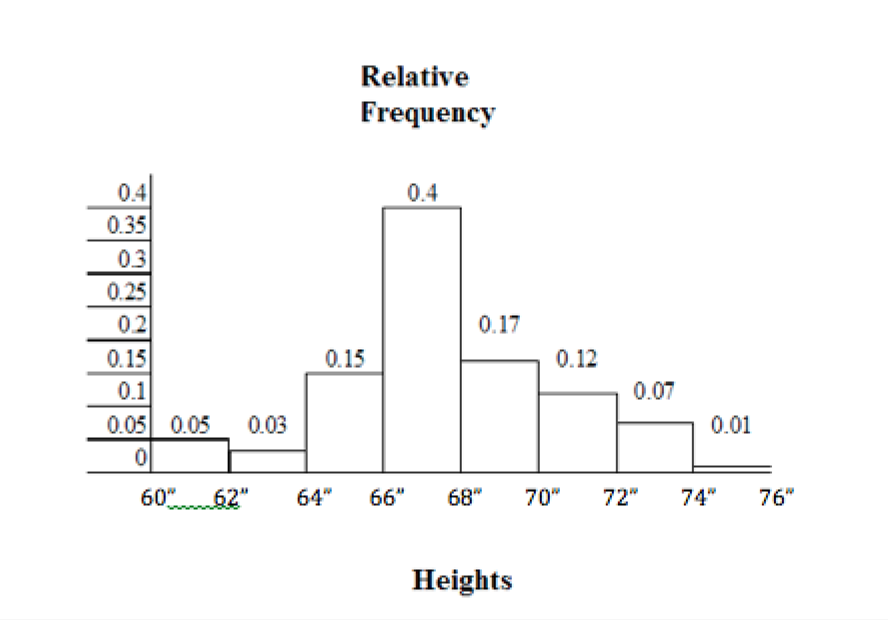
The boundaries are:

- 60
- 60 + 2 = 62

- 62 + 2 = 64
- 64 + 2 = 66
- 66 + 2 = 68
- 68 + 2 = 70
- 70 + 2 = 72
- 72 + 2 = 74
- 74 + 2 = 76

The heights 60 through 61.5 inches are in the interval 60 - 62. The heights that are 63.5 are in the interval 62 - 64. The heights that are 64 through 64.5 are in the interval 64 - 66. The heights 66 through 67.5 are in the interval 66 - 68. The heights 68 through 69.5 are in the interval 68 - 70. The heights 70 through 71 are in the interval 70 -72. The heights 72 through 73.5 are in the interval 72 - 74. The height 74 is in the interval 74 - 76.

The following histogram displays the heights on the x-axis and relative frequency on the y-axis.



Example:

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data since books are counted.
1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4
5 5 5 5 5
6 6
Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 books.
Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Exercise:

Problem:

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6 and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _______ to _______, the 5 in the middle of the interval from _______ to _______, and the _______ in the middle of the interval from _______ to _______ .

Solution:

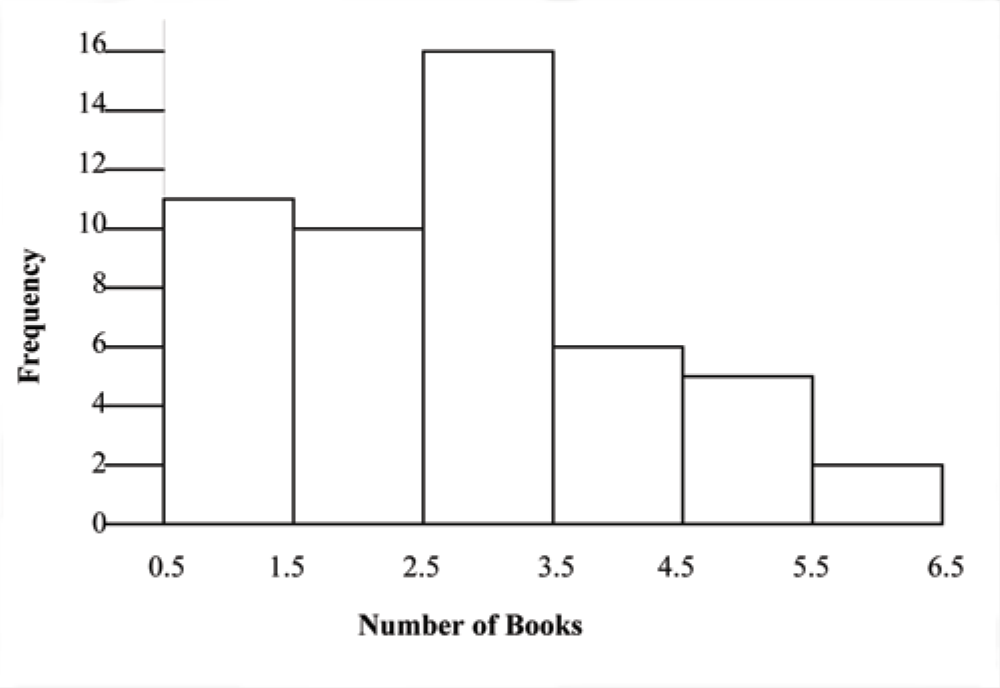
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

Equation:

$$\frac{6.5 - 0.5}{\text{bars}} = 1$$

where 1 is the width of a bar. Therefore, bars $= 6$.
The following histogram displays the number of books on the x-axis and the frequency on the y-axis.



Using the TI-83, 83+, 84, 84+ Calculator Instructions
Go to the Appendix (14:Appendix) in the menu on the left. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example 2.

- Press Y=. Press CLEAR to clear out any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6
- Into L2, enter 11, 10, 16, 6, 5, 2

- Press WINDOW. Make Xmin = .5, Xmax = 6.5, Xscl = (6.5 - .5)/6, Ymin = -1, Ymax = 20, Yscl = 1, Xres = 1
- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH
- Use the TRACE key and the arrow keys to examine the histogram.

Optional Collaborative Exercise

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals. Discuss, also, the shape of the histogram.

Record the data, in dollars (for example, 1.25 dollars).

Construct a histogram.

Glossary

Frequency
: The number of times a value of the data occurs.

Relative Frequency
: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Descriptive Statistics: Box Plot

Box plots or **box-whisker plots** give a good graphical image of the concentration of the data. They also show how far from most of the data the extreme values are. The box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then again in the section on measuring data in this chapter. We use these values to compare how close other data values are to them.

The **median**, a number, is a way of measuring the "center" of the data. You can think of the median as the "middle value," although it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger. For example, consider the following data:

1 11.5 6 7.2 4 8 9 10 6.8 8.3 2 2 10 1

Ordered from smallest to largest:

1 1 2 2 4 6 **6.8** **7.2** 8 8.3 9 10 10 11.5

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2.
Equation:

$$\frac{6.8+7.2}{2}=7$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile is the middle value of the lower half of the data and the third quartile is the middle value of the upper half of the data. To get the idea, consider the same data set shown above:

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The median or **second quartile** is 7. The lower half of the data is 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.

1 1 2 **2** 4 6 6.8

The number 2, which is part of the data, is the **first quartile**. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is 9.

7.2 8 8.3 **9** 10 10 11.5

The number 9, which is part of the data, is the **third quartile**. Three-fourths of the values are less than 9 and one-fourth of the values are more than 9.

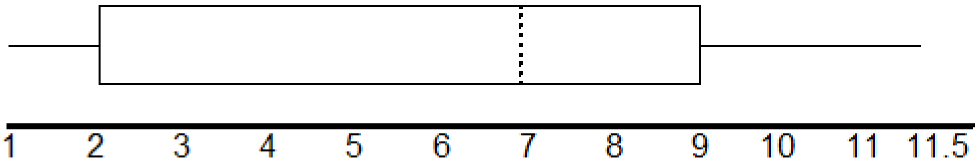
To construct a box plot, use a horizontal number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. **The middle fifty percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The box plot gives a good quick picture of the data.

Note: You may encounter box and whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider the following data:

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The first quartile is 2, the median is 7, and the third quartile is 9. The smallest value is 1 and the largest value is 11.5. The box plot is constructed as follows (see calculator instructions in the back of this book or on the TI web site):



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

Example:
The following data are the heights of 40 students in a statistics class.
59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77
Construct a box plot:
Using the TI-83, 83+, 84, 84+ Calculator

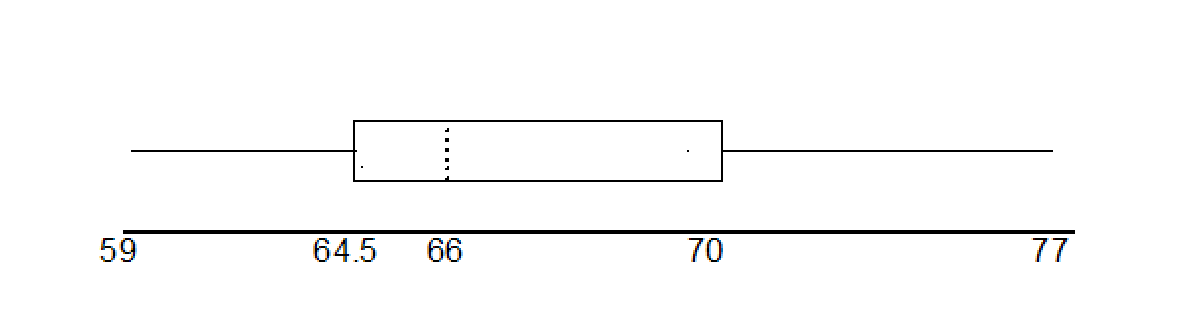
- Enter data into the list editor (Press STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, arrow down.
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.
- Press ENTER
- Use the down and up arrow keys to scroll.

- Smallest value = 59
- Largest value = 77
- Q1: First quartile = 64.5

- Q2: Second quartile or median= 66
- Q3: Third quartile = 70

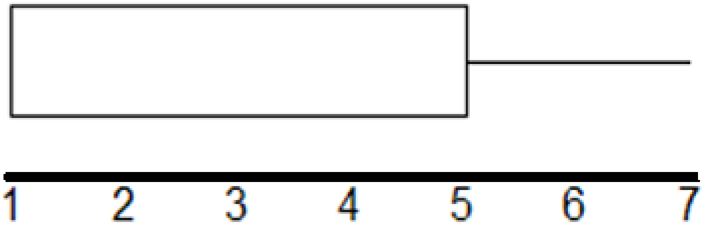
Using the TI-83, 83+, 84, 84+ to Construct the Box Plot
Go to 14:Appendix for Notes for the TI-83, 83+, 84, 84+ Calculator. To create the box plot:

- Press Y=. If there are any equations, press CLEAR to clear them.
- Press 2nd Y=.
- Press 4:Plotsoff. Press ENTER
- Press 2nd Y=
- Press 1:Plot1. Press ENTER.
- Arrow down and then use the right arrow key to go to the 5th picture which is the box plot. Press ENTER.
- Arrow down to Xlist: Press 2nd 1 for L1
- Arrow down to Freq: Press ALPHA. Press 1.
- Press ZOOM. Press 9:ZoomStat.
- Press TRACE and use the arrow keys to examine the box plot.



- **a**Each quarter has 25% of the data.
- **b**The spreads of the four quarters are 64.5 - 59 = 5.5 (first quarter), 66 - 64.5 = 1.5 (second quarter), 70 - 66 = 4 (3rd quarter), and 77 - 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
- **c**Interquartile Range: $\text{IQR} = \text{Q3} - \text{Q1} = 70 - 64.5 = 5.5$.
- **d**The interval 59 through 65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
- **e**The middle 50% (middle half) of the data has a range of 5.5 inches.

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the largest value was 7, the box plot would look as follows:



Example:

Test scores for a college statistics class held during the day are:
99 56 78 55.5 32 90 80 81 56 59 45 77 84.5 84 70 72 68 32 79 90
Test scores for a college statistics class held during the evening are:
98 78 68 83 81 89 88 76 65 45 98 90 80 84.5 85 79 78 98 90 79 81 25.5

Exercise:

Problem:

- What are the smallest and largest data values for each data set?
- What is the median, the first quartile, and the third quartile for each data set?
- Create a boxplot for each set of data.
- Which boxplot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

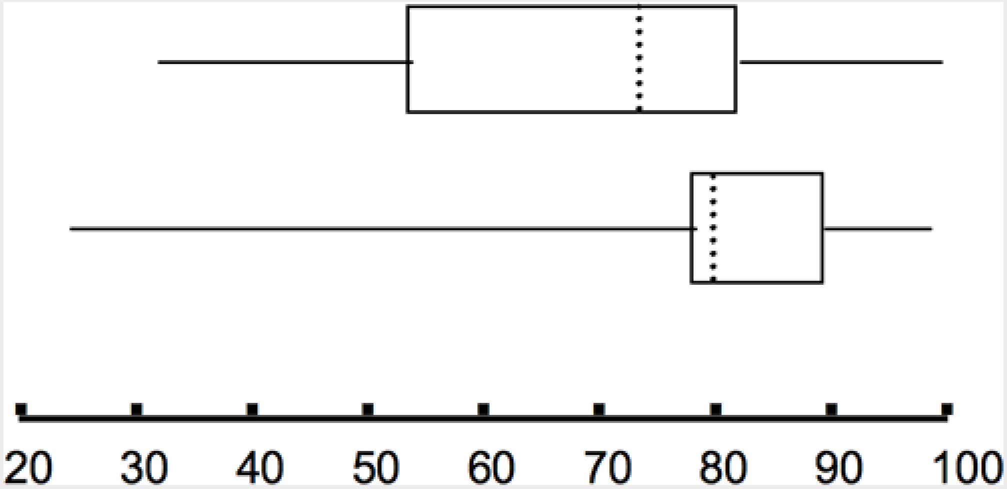
- For each data set, what percent of the data is between the smallest value and the first quartile? (Answer: 25%) the first quartile and the median? (Answer: 25%) the median and the third quartile? the third quartile and the largest value? What percent of the data is between the first quartile and the largest value? (Answer: 75%)

Solution:
First Data Set

- Xmin $= 32$
- Q1 $= 56$
- $M = 74.5$
- Q3 $= 82.5$
- Xmax $= 99$

Second Data Set

- Xmin $= 25.5$
- Q1 $= 78$
- $M = 81$
- Q3 $= 89$
- Xmax $= 98$



The first data set (the top box plot) has the widest spread for the middle 50% of the data. $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$ is $82.5 - 56 = 26.5$ for the first data set and $89 - 78 = 11$ for the second data set. So, the first set of data has its middle 50% of scores more spread out.
25% of the data is between M and $\mathrm{Q3}$ and 25% is between $\mathrm{Q3}$ and Xmax.

Glossary

Median

A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Quartiles

The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Descriptive Statistics: Measuring the Location of the Data
Descriptive Statistics: Measuring the Location of Data explains percentiles and quartiles and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom contributed the section "Interpreting Percentiles, Quartile and the Median."

The common measures of location are **quartiles** and **percentiles** (%iles). Quartiles are special percentiles. The first quartile, Q_1 is the same as the 25th percentile (25th %ile) and the third quartile, Q_3, is the same as the 75th percentile (75th %ile). The median, M, is called both the second quartile and the 50th percentile (50th %ile).

Note: Quartiles are given special attention in the Box Plots module in this chapter.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Recall that quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).
Equation:

$$\text{IQR} = Q_3 - Q_1$$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than** $(1.5)(\text{IQR})$ **below the first quartile or more than** $(1.5)(\text{IQR})$ **above the third quartile**. Potential outliers always need further investigation.

Example:
Exercise:

Problem:

For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

389,950 230,500 158,000 479,000 639,000 114,950 5,500,000 387,000 659,000 529,000 575,000 488,800 1,095,000

Solution:

Order the data from smallest to largest.

114,950 158,000 230,500 387,000 389,950 479,000 488,800 529,000 575,000 639,000 659,000 1,095,000 5,500,000

$M = 488{,}800$

$Q_1 = \frac{230500+387000}{2} = 308750$

$Q_3 = \frac{639000+659000}{2} = 649000$

$\text{IQR} = 649000 - 308750 = 340250$

$(1.5)(\text{IQR}) = (1.5)(340250) = 510375$

$Q_1 - (1.5)(\text{IQR}) = 308750 - 510375 = -201625$

$Q_3 + (1.5)(\text{IQR}) = 649000 + 510375 = 1159375$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential outlier.

Example:
Exercise:

Problem:

For the two data sets in the test scores example, find the following:

- **a**The interquartile range. Compare the two interquartile ranges.
- **b**Any outliers in either set.
- **c**The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

Solution:

For the IQRs, see the answer to the test scores example. The first data set has the larger IQR, so the scores between Q3 and Q1 (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \ \cdot \ \left(\text{IQR}\right) \ = \ \left(\frac{3}{2}\right) \ \cdot \ \left(26.5\right) \ = \ 39.75$
- $\text{Xmax} \ \text{-} \ \text{Q3} \ = \ 99 \ \text{-} \ 82.5 \ = \ 16.5$
- $\text{Q1} \ \text{-} \ \text{Xmin} \ = \ 56 \ \text{-} \ 32 \ = \ 24$

$\left(\frac{3}{2}\right) \cdot \left(\text{IQR}\right) = 39.75$ is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot \left(\text{IQR}\right) = \left(\frac{3}{2}\right) \cdot \left(11\right) = 16.5$
- $\text{Xmax} - \text{Q3} = 98 - 89 = 9$
- $\text{Q1} - \text{Xmin} = 78 - 25.5 = 52.5$

$\left(\frac{3}{2}\right) \cdot \left(\text{IQR}\right) = 16.5$ is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see "Frequency" from the Sampling and

Data Chapter). Get the percentiles from that chart.

First Data Set

- $30\text{th }\%\text{ile (between the 6th and 7th values)} = \frac{(56+59)}{2} = 57.5$
- $80\text{th }\%\text{ile (between the 16th and 17th values)} = \frac{(84+84.5)}{2} = 84.25$

Second Data Set

- $30\text{th }\%\text{ile (7th value)} = 78$
- $80\text{th }\%\text{ile (18th value)} = 90$

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Example:
Finding Quartiles and Percentiles Using a Table

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**

Find the median: Look again at the "cumulative relative frequency " column and find 0.52. The median is the 50th %ile or the second quartile. 50% of 50 = 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th %ile is between the 25th (7) and 26th (7) values. **The median is 7.**

Find the third quartile: The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8** . Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3, is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example:

Exercise:

Problem: Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile.
4. What is another name for the first quartile?

Solution:

1. $\frac{(8+9)}{2} = 8.5$
   Look where cum. rel. freq. = 0.80. 80% of the data is 8 or less. 80th %ile is between the last 8 and first 9.
2. 9
3. 6
4. First Quartile = 25th %ile

Collaborative Classroom Exercise: Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions.

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct a table of the data.
4. Construct 2 different histograms. For each, starting value = ______ ending value = _____.
5. Use the table to find the median, first quartile, and third quartile.
6. Construct a box plot.
7. Use the table to find the following:
   - The 10th percentile
   - The 70th percentile
   - The percent of students who own less than 4 sweaters

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order, from smallest to largest. p% of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad". The interpretation of whether a certain percentile is good or bad depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good'; in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to properly interpret percentiles is important not only when describing data, but is also important in later chapters of this textbook when calculating probabilities.

Guideline:

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information:

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example:
On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example:
On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile could be considered good, as answering more questions correctly is desirable.

Example:
At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units
- 70% of students are enrolled in 7 or more credit units
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Do the following Practice Problems for Interpreting Percentiles
Exercise:

Problem:

- **a** For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
- **b** The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.

- **c** A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Solution:

- **a** For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster.
- **b**INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer.
- **c**He is among the slowest cyclists (90% of cyclists were faster than him.) INTERPRETATION: 90% of cyclists had a finish time of 1 hour, 12 minutes or less.Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer

Exercise:

Problem:

- **a** For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
- **b**The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

- **a**For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed, which is faster.
- **b**INTERPRETATION: 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise:

Problem:

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Solution:

On an exam you would prefer a high percentile; higher percentiles correspond to higher grades on the exam.

Exercise:

Problem:

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution:

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than you did. In this context, you would prefer a wait time corresponding to a lower percentile. INTERPRETATION: 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise:

Problem:

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Solution:

Li should be pleased. Her salary is relatively high compared to other recent college grads. 78% of recent college graduates earn less than Li does. 22% of recent college graduates earn more than Li does.

Exercise:

Problem:

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Solution:

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise:

Problem:

- The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:
- a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
- b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percent of students from each high school are "eligible in the local context"?

Solution:

- **a**The top 12% of students are those who are at or above the **88th percentile** of admissions index scores.
- **b** The **top 4%** of students' GPAs are at or above the 96th percentile, making the top 4% of students "eligible in the local context".

Exercise:

Problem:

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution:

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

**With contributions from Roberta Bloom

Glossary

Interquartile Range (IRQ)
: The distance between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 - Q1.

Outlier
: An observation that does not fit the rest of the data.

Percentile
: A number that divides ordered data into hundredths.

Example:
Let a data set contain 200 ordered observations starting with $\{2.3, 2.7, 2.8, 2.9, 2.9, 3.0...\}$. Then the first percentile is $\frac{(2.7+2.8)}{2} = 2.75$, because 1% of the data is to the left of this point on the number line and 99% of the data is on its right. The second percentile is $\frac{(2.9+2.9)}{2} = 2.9$. Percentiles may or may not be part of the data. In this example, the first percentile is not in the data, but the second percentile is. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Descriptive Statistics: Measuring the Center of the Data

This chapter discusses measuring descriptive statistical information using the center of the data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts (previously discussed under box plots in this chapter). The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note: The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

The mean can also be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced "x bar"): x.

The Greek letter μ (pronounced "mew") represents the population mean. One of the requirements for the sample mean to be a good estimate of the population mean is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

11122344444

Equation:

$$x = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

Equation:

$$x = \frac{3 \times 1 + 2 \times 2{+}1 \times 3 + 5 \times 4}{11} = 2.7$$

In the second calculation for the sample mean, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by 2 after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2}$= $\frac{97+1}{2}$= 49. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2}$= $\frac{100+1}{2}$= 50.5. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example:
Exercise:

Problem:

AIDS data indicating the number of months an AIDS patient lives after taking a new antibody drug are as follows (smallest to largest):

34881011121314151516161717182122222424252626272729293132333
33434353740444447

Calculate the mean and the median.

Solution:

The calculation for the mean is:

$$x = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, **M**, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

34881011121314151516161717182122222424
25262627272929313233333434353740444447

$$M = \frac{24+24}{2} = 24$$

The median is 24.

Using the TI-83,83+,84, 84+ Calculators
Calculator Instructions are located in the menu item 14:Appendix (Notes for the TI-83, 83+, 84, 84+ Calculators).

- Enter data into the list editor. Press STAT 1:EDIT
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and ENTER.
- Press the down and up arrow keys to scroll.

$x = 23.6$, $M = 24$

Example:
Exercise:

Problem:

Suppose that, in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center," the mean or the median?

Solution:

$x = \frac{5000000+49\times30000}{50} = 129400$

$M = 30000$

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. If a data set has two values that occur the same number of times, then the set is bimodal.

Example:
Statistics exam scores for 20 students are as follows
Statistics exam scores for 20 students are as follows:
50 53 59 59 63 63 72 72 72 72 72 76 78 81 83 84 84 84 90 93
Exercise:

Problem: Find the mode.

Solution:

The most frequent score is 72, which occurs five times. Mode = 72.

Example:
Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

Note: The mode can be calculated for qualitative data as well as for quantitative data.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean x of the sample is very likely to get closer and closer to μ. This is discussed in more detail in **The Central Limit Theorem**.

Note: The formula for the mean is located in the Summary of Formulas section course.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	5/30
1	15/30
2	6/30
3	4/30
4	1/30

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean x is an example of a statistic which estimates the population mean μ.

Glossary

Mean

A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x) is $x = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median

A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Mode

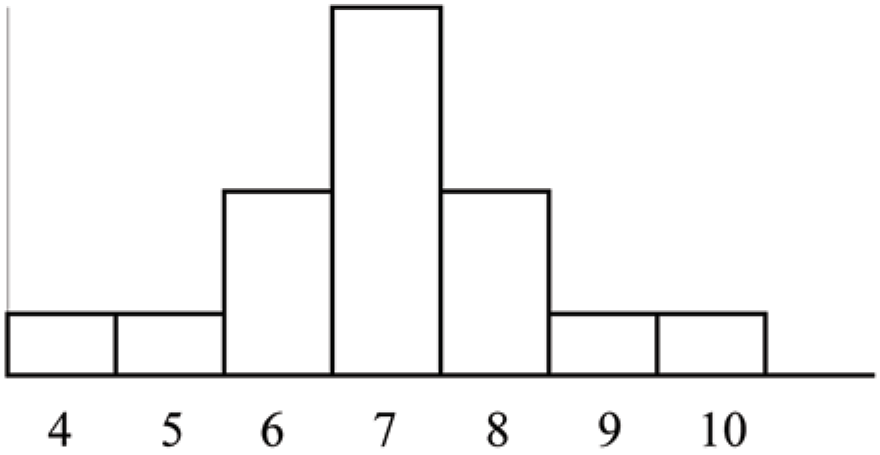
The value that appears most frequently in a set of data.

Descriptive Statistics: Skewness and the Mean, Median, and Mode

Consider the following data set:

4 5 6 6 6 7 7 7 7 7 7 8 8 8 9 10

This data set produces the histogram shown below. Each interval has width one and each value is located in the middle of an interval.

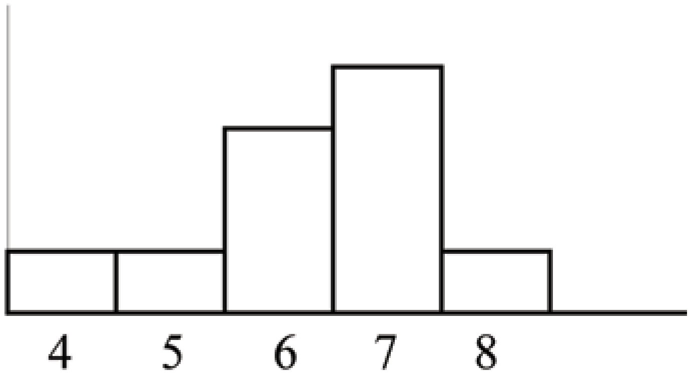


The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal) and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data:

4 5 6 6 6 7 7 7 7 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. The shape distribution is called **skewed to the left** because it is pulled out to the left.

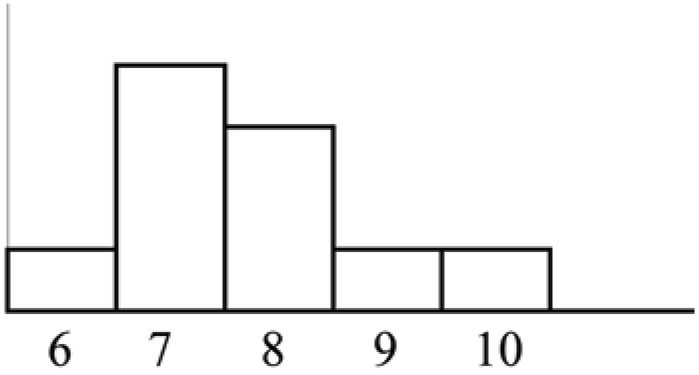


The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

The histogram for the data:

6 7 7 7 7 8 8 8 9 10

is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Descriptive Statistics: Measuring the Spread of the Data
Descriptive Statistics: Measuring the Spread of Data explains standard deviation as a measure of variation in data and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom made contributions that helped to clarify the standard deviation and the variance.

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation.

The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set
- can be used to determine whether a particular data value is close to or far from the mean

The standard deviation provides a measure of the overall variation in a data set
The standard deviation is always positive or 0. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying waiting times at the checkout line for customers at supermarket A and supermarket B; the average wait time at both markets is 5 minutes. At market A, the standard deviation for the waiting time is 2 minutes; at market B the standard deviation for the waiting time is 4 minutes.

Because market B has a higher standard deviation, we know that there is more variation in the waiting times at market B. Overall, wait times at market B are more spread out from the average; wait times at market A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.
Suppose that Rosa and Binh both shop at Market A. Rosa waits for 7 minutes and Binh waits for 1 minute at the checkout counter. At market A, the mean wait time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

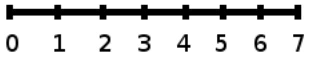
- 7 is 2 minutes longer than the average of 5; 2 minutes is equal to one standard deviation.
- Rosa's wait time of 7 minutes is **2 minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.

Binh waits for 1 minute.

- 1 is 4 minutes less than the average of 5; 4 minutes is equal to two standard deviations.
- Binh's wait time of 1 minute is **4 minutes less than the average** of 5 minutes.
- Binh's wait time of 1 minute is **two standard deviations below the average** of 5 minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than 2 standard deviations. (We will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that 7 is **one** standard deviation to the **right** of 5 because
$5 + (1)(2) = 7$.

If 1 were also part of the data set, then 1 is **two** standard deviations to the **left** of 5 because
$5 + (\text{-}2)(2) = 1$.



- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- 7 is **one standard deviation more than the mean** of 5 because: 7=5+**(1)**(2)
- 1 is **two standard deviations less than the mean** of 5 because: 1=5+**(−2)**(2)

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population:

- **sample:** $x = x + (\#\text{ofSTDEV})(\text{s})$
- **Population:** $x = \mu + (\#\text{ofSTDEV})(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol x is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation
If x is a number, then the difference "x - mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - x$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is an **average of the squares of the deviations** (the $x - x$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by **N**, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by **n-1**, one less than the number of items in the sample. You can see that in the formulas below.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f\cdot(x-x)^2}{n-1}}$
- For the sample standard deviation, the denominator is **n-1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f\cdot(x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is **N**, the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is 1. If a value appears three times in the data set or population, f is 3.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in **Descriptive Statistics: Measuring the Center of the Data**. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in **The Central Limit Theorem** (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note: In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83,83+,84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean.

Example:

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9 9.5 9.5 10 10 10 10 10.5 10.5 10.5 10.5 11 11 11 11 11 11 11.5 11.5 11.5

Equation:

$$x = \frac{9 + 9.5 \times 2 + 10 \times 4 + 10.5 \times 4 + 11 \times 6 + 11.5 \times 3}{20} = 10.525$$

The average age is 10.53 years, rounded to 2 places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	**Freq.**	**Deviations**	Deviations^2	**(Freq.)**(Deviations^2**)**
x	f	$(x - x)$	$(x - x)^2$	$(f)(x - x)^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times .275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times .000625 = .0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times .225625 = 1.35375$

Data	**Freq.**	**Deviations**	Deviations^2	**(Freq.)**(Deviations^2**)**
11.5	*3*	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times .950625 = 2.851875$

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 - 1):
$s^2 = \frac{9.7375}{20-1} = 0.5125$
The **sample standard deviation** s is equal to the square root of the sample variance:
$s = \sqrt{0.5125} =. 0715891$ Rounded to two decimal places, $s = 0.72$
Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.
Exercise:

Problem: Verify the mean and standard deviation calculated above on your calculator or computer.

Solution:
Using the TI-83,83+,84+ Calculators

- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- x=10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**
- For a sample: x = x + (#ofSTDEVs)(s)
- For a population: x = μ + (#ofSTDEVs)(σ)
- For this example, use x = x + (#ofSTDEVs)(s) because the data is from a sample

Exercise:

Problem: Find the value that is 1 standard deviation above the mean. Find $(x + 1s)$.

Solution:

$(x + 1s) = 10.53 + (1)(0.72) = 11.25$

Exercise:

Problem: Find the value that is two standard deviations below the mean. Find $(x - 2s)$.

Solution:

$(x - 2s) = 10.53 - (2)(0.72) = 9.09$

Exercise:

Problem: Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

- $(x - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
- $(x + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Explanation of the standard deviation calculation shown in the table
The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11. The deviations 0.97 and 0.47 indicate that. A positive deviation occurs when the data value is greater than the mean. A negative deviation occurs when the data value is less than the mean; the deviation is -1.525 for the data value 9. **If you add the deviations, the sum is always zero**. (For this example, there are n=20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by n=20, the calculation divided by n-1=20-1=19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one (n-1). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by (n-1) gives a better estimate of the population variance.

Note: Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is 0, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**.

Note: The formula for the standard deviation is at the end of the chapter.

Example:
Exercise:

Problem: Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

3342494953555561 6367686869697273 7478808388888890 929494949496100

- **a**Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- **b**Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
  - **i**The sample mean
  - **ii**The sample standard deviation
  - **iii**The median
  - **iv**The first quartile
  - **v**The third quartile
  - **vi**IQR
- **c**Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

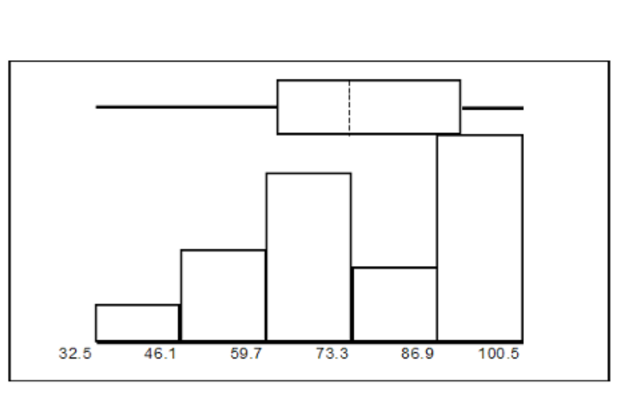
Solution:

- **a**

Data	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	**0.998** (Why isn't this value 1?)

- **b**
  - **i**The sample mean = 73.5
  - **ii**The sample standard deviation = 17.9
  - **iii**The median = 73
  - **iv**The first quartile = 61
  - **v**The third quartile = 90
  - **vi**IQR = 90 - 61 = 29
- **c**The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram; number of intervals is 5 for the histogram so the width of an interval is (100.5 - 32.5) divided by 5 which is equal to 13.6. Endpoints of the intervals: starting point is 32.5, 32.5+13.6 = 46.1, 46.1+13.6 = 59.7, 59.7+13.6 = 73.3, 73.3+13.6 = 86.9, 86.9+13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 - 33 = 40) than the spread in the upper 50% (100 - 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (IQR = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, it can be misleading to compare the data values directly.

- For each data value, calculate how many standard deviations the value is away from its mean.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#\text{ofSTDEVs} = \frac{\text{value}-\text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	x = x + z s	$z = \frac{x-x}{s}$
Population	x = μ + z σ	$z = \frac{x-\mu}{\sigma}$

Example:
Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. when compared to his school?

Student	**GPA**	**School Mean GPA**	**School Standard Deviation**
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$\#\text{ofSTDEVs} = \frac{\text{value}-\text{mean}}{\text{standard deviation}}$; $z = \frac{x-\mu}{\sigma}$

For John, $z = \#\text{ofSTDEVs} = \frac{2.85-3.0}{0.7} = -0.21$

For Ali, $z = \#\text{ofSTDEVs} = \frac{77-80}{10} = -0.3$

John has the better G.P.A. when compared to his school because his G.P.A. is 0.21 standard deviations **below** his school's mean while Ali's G.P.A. is 0.3 standard deviations **below** his school's mean.

John's z-score of −0.21 is higher than Ali's z-score of −0.3 . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4 1/2 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is MOUND-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within 1 standard deviation of the mean.
- Approximately 95% of the data is within 2 standard deviations of the mean.
- More than 99% of the data is within 3 standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is mound-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

**With contributions from Roberta Bloom

Glossary

Standard Deviation

> A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variance

> Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - x$ where x is a value of the data and x is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

Descriptive Statistics: Summary of Formulas

A summary of useful formulas used in examining descriptive statistics

Commonly Used Symbols

- The symbol Σ means to add or to find the sum.
- n = the number of data values in a sample
- N = the number of people, things, etc. in the population
- x = the sample mean
- s = the sample standard deviation
- μ = the population mean
- σ = the population standard deviation
- f = frequency
- x = numerical value

Commonly Used Expressions

- $x{*}f$ = A value multiplied by its respective frequency
- $\sum x$ = The sum of the values
- $\sum x{*}f$ = The sum of values multiplied by their respective frequencies
- $(x-x)$ or $(x-\mu)$ = Deviations from the mean (how far a value is from the mean)
- $(x-x)^2$ or $(x-\mu)^2$ = Deviations squared
- $f(x-x)^2$ or $f(x-\mu)^2$ = The deviations squared and multiplied by their frequencies

Mean Formulas:

- $x = \frac{\sum x}{n}$ or $x = \frac{\sum f\cdot x}{n}$
- $\mu = \frac{\sum x}{N}$ or μ= $\frac{\sum f\cdot x}{N}$

Standard Deviation Formulas:

- $s = \sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f\cdot(x-x)^2}{n-1}}$
- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f\cdot(x-\mu)^2}{N}}$

Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- x = x+ (#ofSTDEVs)(s)
- x = μ + (#ofSTDEVs)(σ)

Descriptive Statistics: Practice 1
This module provides students with opportunities to apply concepts related to descriptive statistics. Students are asked to take a set of sample data and calculate a series of statistical values for that data.

Student Learning Outcomes

- The student will calculate and interpret the center, spread, and location of the data.
- The student will construct and interpret histograms an box plots.

Given

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Complete the Table

Data Value (# cars)	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**

Discussion Questions

Exercise:

Problem: What does the frequency column sum to? Why?

Solution:

65

Exercise:

Problem: What does the relative frequency column sum to? Why?

Solution:

1

Exercise:

Problem:

What is the difference between relative frequency and frequency for each data value?

Exercise:

Problem:

What is the difference between cumulative relative frequency and relative frequency for each data value?

Enter the Data

Enter your data into your calculator or computer.

Construct a Histogram

Determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram below. Label the horizontal and vertical axes with words. Include numerical scaling.

Data Statistics

Calculate the following values:

Exercise:

Problem: Sample mean = x =

Solution:

4.75

Exercise:

Problem: Sample standard deviation = s_x =

Solution:

1.39

Exercise:

Problem: Sample size = n =

Solution:

65

Calculations

Use the table in section 2.11.3 to calculate the following values:
Exercise:

Problem: Median =

Solution:

4

Exercise:

Problem: Mode =

Solution:

4

Exercise:

Problem: First quartile =

Solution:

4

Exercise:

Problem: Second quartile = median = 50th percentile =

Solution:

4

Exercise:

Problem: Third quartile =

Solution:

6

Exercise:

Problem: Interquartile range (IQR) = _____ - _____ = _____

Solution:

$6 - 4 = 2$

Exercise:

Problem: 10th percentile =

Solution:

3

Exercise:

Problem: 70th percentile =

Solution:

6

Exercise:

Problem: Find the value that is 3 standard deviations:

- **a**Above the mean
- **b**Below the mean

Solution:

- **a**8.93

- **b**0.58

Box Plot

Construct a box plot below. Use a ruler to measure and scale accurately.

Interpretation

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Descriptive Statistics: Homework
Descriptive Statistics: Homework is part of the collection col10555 written by Barbara Illowsky and Susan Dean and provides homework questions related to lessons about descriptive statistics.

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

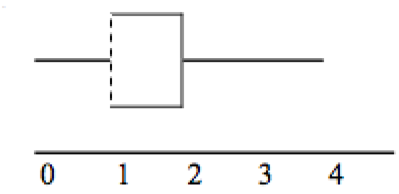
# of movies	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0	5		
1	9		
2	6		
3	4		
4	1		

- **a**Find the sample mean x
- **b**Find the sample standard deviation, s
- **c**Construct a histogram of the data.
- **d**Complete the columns of the chart.
- **e**Find the first quartile.
- **f**Find the median.
- **g**Find the third quartile.
- **h**Construct a box plot of the data.

- **i**What percent of the students saw fewer than three movies?
- **j**Find the 40th percentile.
- **k**Find the 90th percentile.
- **l**Construct a line graph of the data.
- **m**Construct a stem plot of the data.

Solution:

- **a**1.48
- **b**1.12
- **e**1
- **f**1
- **g**2
- **h**



- **i**80%
- **j**1
- **k**3

Exercise:

Problem:

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years. ((*Source: http://www.usatoday.com/news/nation/story/2012-05-17/minority-births-census/55029100/1)*)

- **a**Based upon this information, give two reasons why the black median age could be lower than the white median age.
- **b**Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?

- **c**How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercise:

Problem:

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

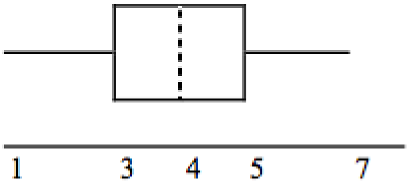
X	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	5		
3	8		
4	12		
5	12		
7	1		

- **a**Find the sample mean x
- **b**Find the sample standard deviation, s
- **c**Construct a histogram of the data.
- **d**Complete the columns of the chart.
- **e**Find the first quartile.

- **f**Find the median.
- **g**Find the third quartile.
- **h**Construct a box plot of the data.
- **i**What percent of the students owned at least five pairs?
- **j**Find the 40th percentile.
- **k**Find the 90th percentile.
- **l**Construct a line graph of the data
- **m**Construct a stem plot of the data

Solution:

- **a**3.78
- **b**1.29
- **e**3
- **f**4
- **g**5
- **h**



- **i**32.5%
- **j**4
- **k**5

Exercise:

Problem:

600 adult Americans were asked by telephone poll, What do you think constitutes a middle-class income? The results are below. Also, include left endpoint, but not the right endpoint. (*Source: Time magazine; survey by Yankelovich Partners, Inc.*)

Note:"Not sure" answers were omitted from the results.

Salary ($)	**Relative Frequency**
< 20,000	0.02
20,000 - 25,000	0.09
25,000 - 30,000	0.19
30,000 - 40,000	0.26
40,000 - 50,000	0.18
50,000 - 75,000	0.17
75,000 - 99,999	0.02
100,000+	0.01

- **a**What percent of the survey answered "not sure" ?
- **b**What percent think that middle-class is from $25,000 - $50,000 ?
- **c**Construct a histogram of the data

    1. **i**Should all bars have the same width, based on the data? Why or why not?
    2. **ii**How should the <20,000 and the 100,000+ intervals be handled? Why?

- **d**Find the 40th and 80th percentiles
- **e**Construct a bar graph of the data

Exercise:

Problem:

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year (*Source: San Jose Mercury News*)

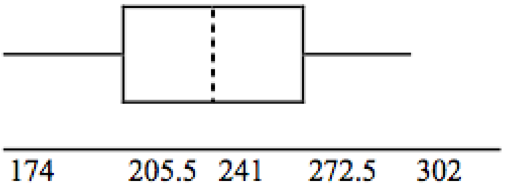
177 205 210 210 232 205 185 185 178 210 206 212 184 174 185 242 188 212 215 247 241 223 220 260 245 259 278 270 280 295 275 285 290 272 273 280 285 286 200 215 185 230 250 241 190 260 250 302 265 290 276 228 265

- **a**Organize the data from smallest to largest value.
- **b**Find the median.
- **c**Find the first quartile.
- **d**Find the third quartile.
- **e**Construct a box plot of the data.
- **f**The middle 50% of the weights are from _______ to _______.
- **g**If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
- **h**If our population were the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
- **i**Assume the population was the San Francisco 49ers. Find:
  - **i**the population mean, μ.
  - **ii**the population standard deviation, σ.
  - **iii**the weight that is 2 standard deviations below the mean.
  - **iv**When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
- **j**That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who

was lighter, Smith or Young? How did you determine your answer?

Solution:

- **b**241
- **c**205.5
- **d**272.5
- **e**



- **f**205.5, 272.5
- **g**sample
- **h**population
- **i**
  - **i**236.34
  - **ii**37.50
  - **iii**161.34
  - **iv**0.84 std. dev. below the mean
- **j**Young

Exercise:

Problem:

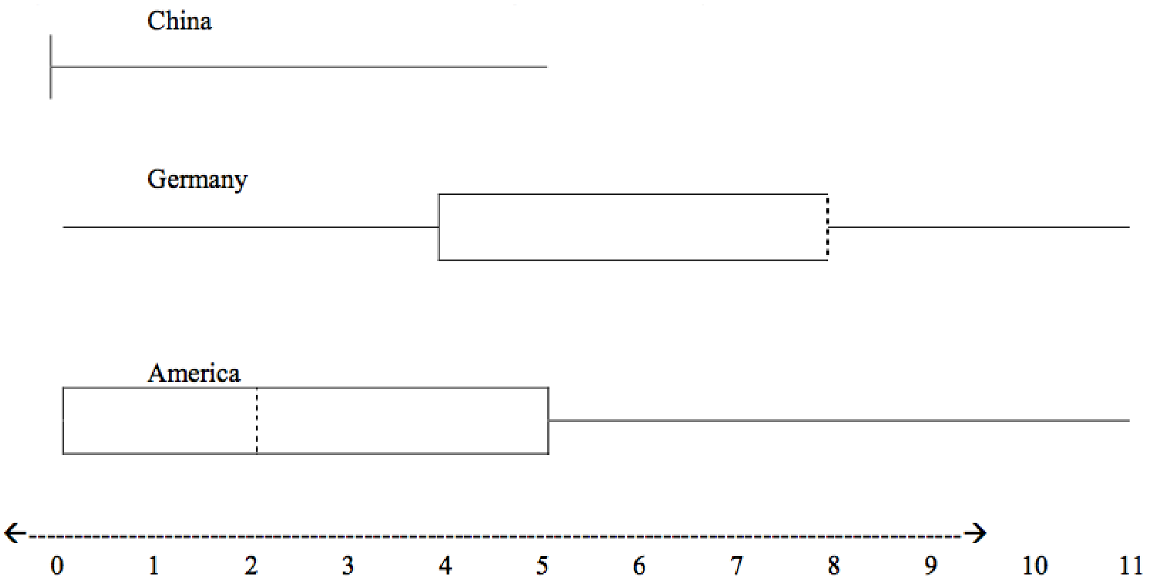
An elementary school class ran 1 mile with a mean of 11 minutes and a standard deviation of 3 minutes. Rachel, a student in the class, ran 1 mile in 8 minutes. A junior high school class ran 1 mile with a mean of 9 minutes and a standard deviation of 2 minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran 1 mile with a mean of 7 minutes and a standard deviation of 4 minutes. Nedda, a student in the class, ran 1 mile in 8 minutes.

- **a**Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
- **b**Who is the fastest runner with respect to his or her class? Explain why.

Exercise:

Problem:

In a survey of 20 year olds in China, Germany and America, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



- **a**In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
- **b**Explain how it is possible that more Americans than Germans surveyed have been to over eight foreign countries.
- **c**Compare the three box plots. What do they imply about the foreign travel of twenty year old residents of the three countries when compared to each other?

Exercise:

Problem:

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The twelve change scores are as follows:

3 8 -1 2 0 5 -3 1 -1 6 5 -2

- **a**What is the mean change score?
- **b**What is the standard deviation for this population?
- **c**What is the median change score?
- **d**Find the change score that is 2.2 standard deviations below the mean.

Exercise:

Problem:

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best G.P.A. when compared to his school? Explain how you determined your answer.

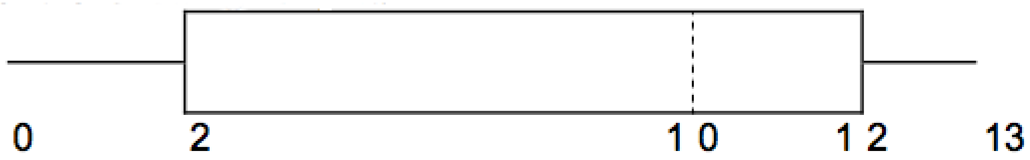
Student	G.P.A.	School Ave. G.P.A.	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Solution:

Kamala

Exercise:

Problem: Given the following box plot:

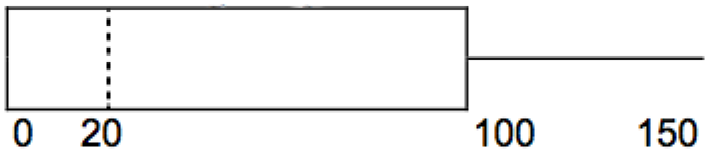


- **a**Which quarter has the smallest spread of data? What is that spread?
- **b**Which quarter has the largest spread of data? What is that spread?
- **c**Find the Inter Quartile Range (IQR).
- **d**Are there more data in the interval 5 - 10 or in the interval 10 - 13? How do you know this?
- **e**Which interval has the fewest data in it? How do you know this?

  - **I** 0-2
  - **II**2-4
  - **III**10-12
  - **IV**12-13

Exercise:

Problem: Given the following box plot:



- **a**Think of an example (in words) where the data might fit into the above box plot. In 2-5 sentences, write down the example.
- **b**What does it mean to have the first and second quartiles so close together, while the second to fourth quartiles are far apart?

Exercise:

Problem:

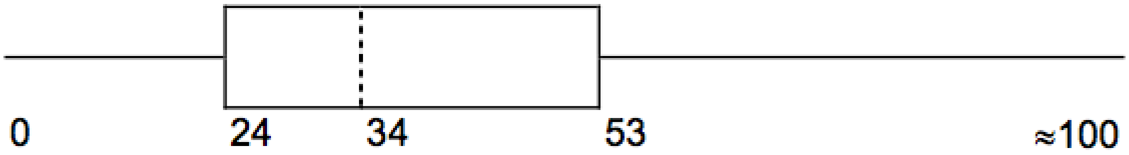
Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows. (*Source: West magazine*)

Age Group	**Percent of Community**
0-17	18.9
18-24	8.0
25-34	22.8
35-44	15.0
45-54	13.1

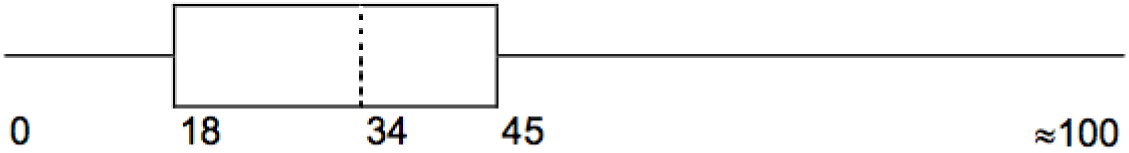
Age Group	**Percent of Community**
55-64	11.9
65+	10.3

- **a**Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not?
- **b**What percent of the community is under age 35?
- **c**Which box plot most resembles the information above?

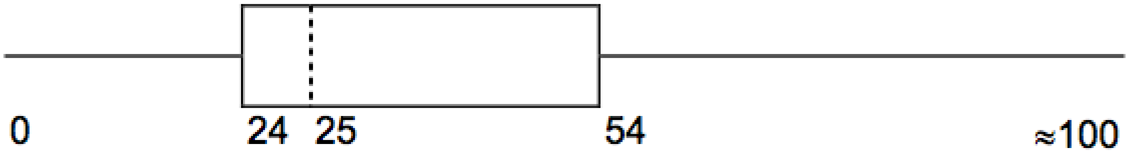
i.



ii.



iii.



Exercise:

Problem:

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, each asked adult consumers the number of fiction paperbacks they had purchased the previous month. The results are below.

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	

# of books	Freq.	Rel. Freq.
5	10	
7	5	
9	1	

Publisher B

# of books	Freq.	Rel. Freq.
0-1	20	
2-3	35	
4-5	12	
6-7	2	
8-9	1	

Publisher C

- **a**Find the relative frequencies for each survey. Write them in the charts.
- **b**Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of 1. For Publisher C, make bar widths of 2.
- **c**In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.

- **d**Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
- **e**Make new histograms for Publisher A and Publisher B. This time, make bar widths of 2.
- **f**Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercise:

Problem:

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all on-board transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Below is a summary of the bills for each group.

Amount($)	**Frequency**	**Rel. Frequency**
51-100	5	
101-150	10	
151-200	15	
201-250	15	
251-300	10	
301-350	5	

Singles

Amount($)	**Frequency**	**Rel. Frequency**
100-150	5	
201-250	5	
251-300	5	
301-350	5	
351-400	10	
401-450	10	
451-500	10	
501-550	10	
551-600	5	
601-650	5	

Couples

- **a**Fill in the relative frequency for each group.
- **b**Construct a histogram for the Singles group. Scale the x-axis by $50. widths. Use relative frequency on the y-axis.
- **c**Construct a histogram for the Couples group. Scale the x-axis by $50. Use relative frequency on the y-axis.
- **d**Compare the two graphs:

  - **i**List two similarities between the graphs.
  - **ii**List two differences between the graphs.
  - **iii**Overall, are the graphs more similar or different?

- **e**Construct a new graph for the Couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by $50, scale it by $100. Use relative frequency on the y-axis.
- **f**Compare the graph for the Singles with the new graph for the Couples:

  - **i**List two similarities between the graphs.
  - **ii**Overall, are the graphs more similar or different?

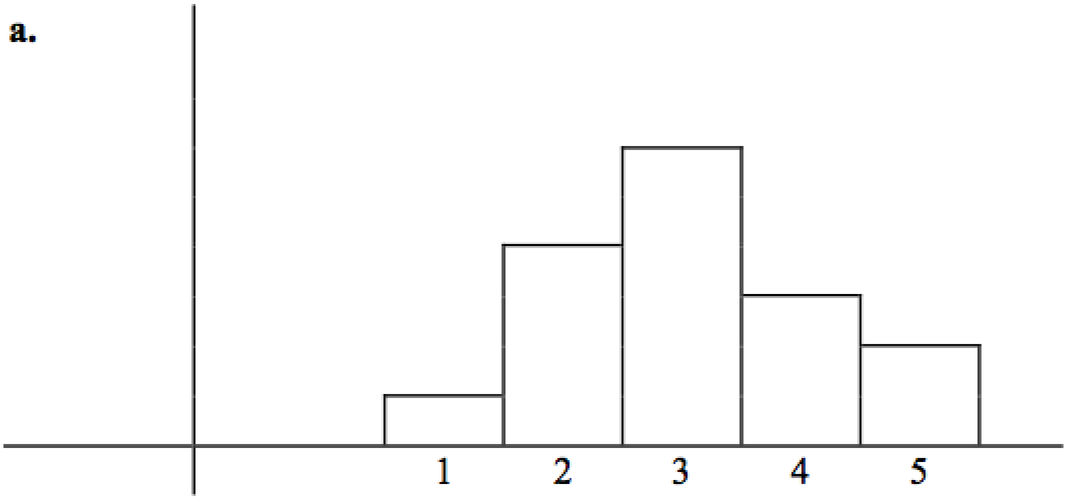
- **i**By scaling the Couples graph differently, how did it change the way you compared it to the Singles?
- **j**Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person in a couple? Explain why in one or two complete sentences.

Exercise:

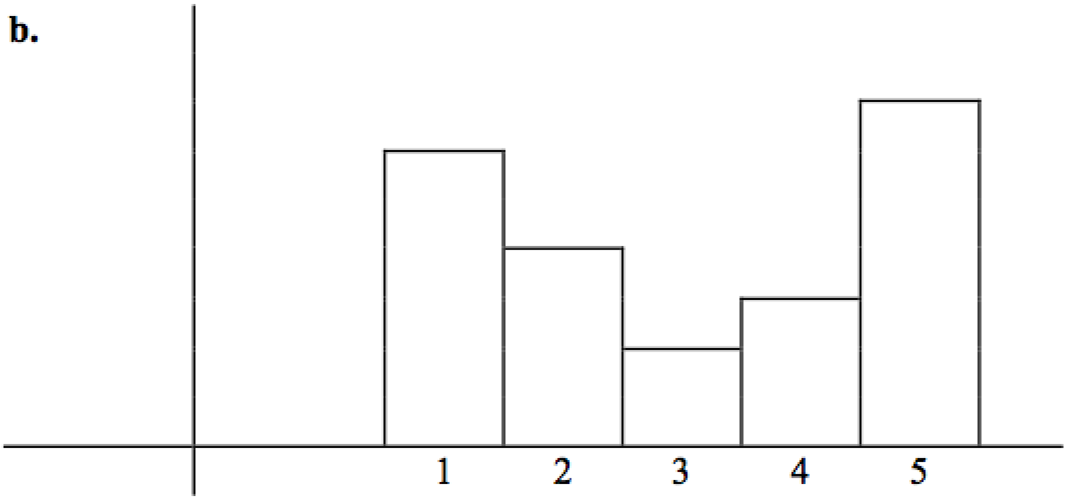
Problem:

Refer to the following histograms and box plot. Determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

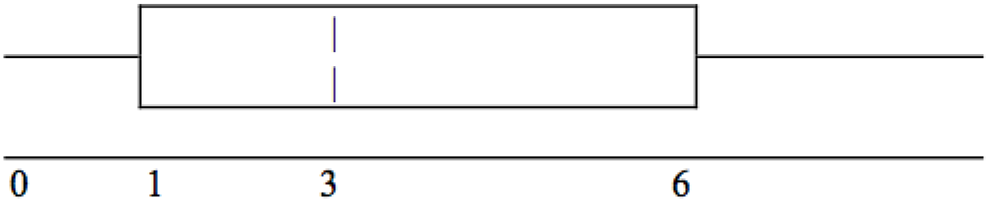
a.



b.



c.



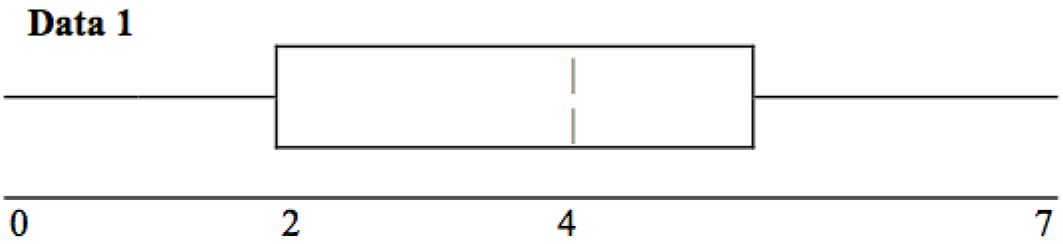
- **a**The medians for all three graphs are the same.
- **b**We cannot determine if any of the means for the three graphs is different.
- **c**The standard deviation for (b) is larger than the standard deviation for (a).
- **d**We cannot determine if any of the third quartiles for the three graphs is different.

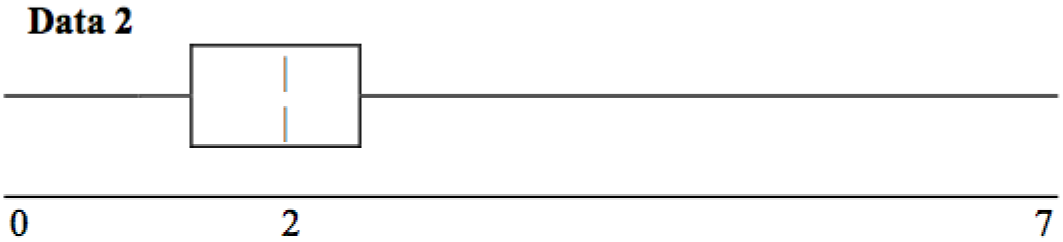
Solution:

- **a**True
- **b**True
- **c**True
- **d**False

Exercise:

Problem: Refer to the following box plots.





- **a**In complete sentences, explain why each statement is false.
  - **i****Data 1** has more data values above 2 than **Data 2** has above 2.
  - **ii**The data sets cannot have the same mode.
  - **iii**For **Data 1**, there are more data values below 4 than there are above 4.
- **b**For which group, Data 1 or Data 2, is the value of “7” more likely to be an outlier? Explain why in complete sentences

Exercise:

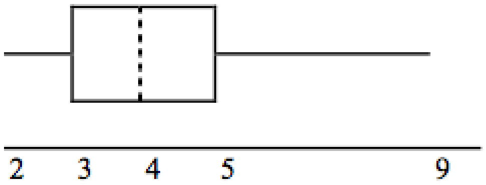
Problem:

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

- **a**Organize the data in a chart.
- **b**Find the median, the first quartile, and the third quartile.
- **c**Find the 65th percentile.
- **d**Find the 10th percentile.
- **e**Construct a box plot of the data.
- **f**The middle 50% of the conferences last from _______ days to _______ days.
- **g**Calculate the sample mean of days of engineering conferences.
- **h**Calculate the sample standard deviation of days of engineering conferences.
- **i**Find the mode.
- **j**If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
- **k**Give two reasons why you think that 3 - 5 days seem to be popular lengths of engineering conferences.

Solution:

- **b**4,3,5
- **c**4
- **d**3
- **e**



- **f**3,5
- **g**3.94
- **h**1.28
- **i**3
- **j**mode

Exercise:

Problem:

A survey of enrollment at 35 community colleges across the United States yielded the following figures (*source: Microsoft Bookshelf*):

6414 1550 2109 9350 21828 4300 5944 5722 2825 2044 5481 5200 5853 2750 10012 6357 27000 9414 7681 3200 17500 9200 7380 18314 6557 13713 17768 7493 2771 2861 1263 7285 28165 5080 11622

- **a**Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
- **b**Construct a histogram of the data.
- **c**If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
- **d**Calculate the sample mean.
- **e**Calculate the sample standard deviation.
- **f**A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Exercise:

Problem:

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years. (*Source: Bureau of the Census*)

- **a**What does it mean for the median age to rise?
- **b**Give two reasons why the median age could rise.
- **c**For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

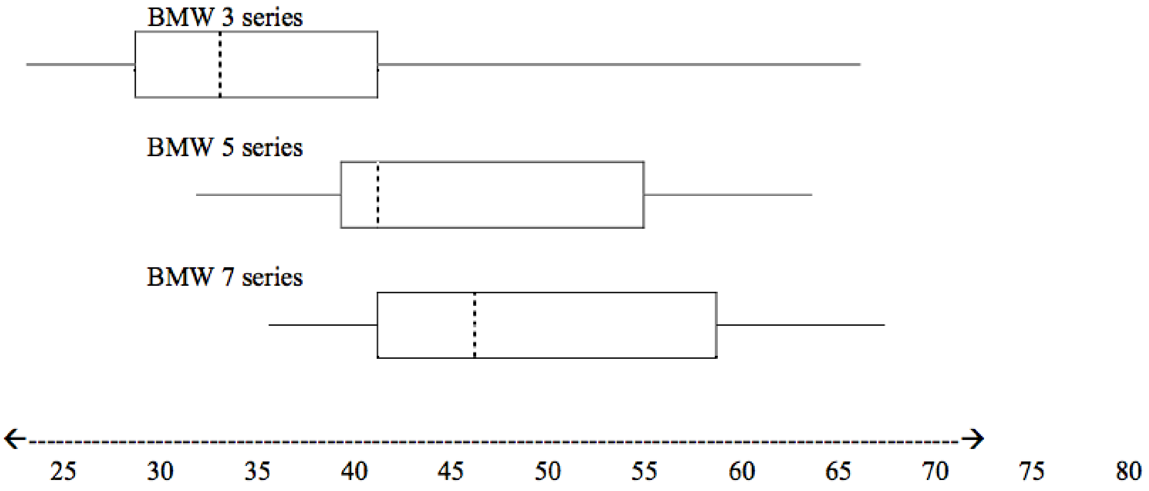
Solution:

- **c**Maybe

Exercise:

Problem:

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



- **a**In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.

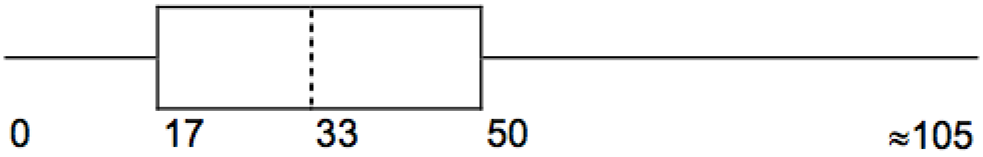
- **b**Which group is most likely to have an outlier? Explain how you determined that.
- **c**Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
- **d**Look at the BMW 5 series. Which quarter has the smallest spread of data? What is that spread?
- **e**Look at the BMW 5 series. Which quarter has the largest spread of data? What is that spread?
- **f**Look at the BMW 5 series. Estimate the Inter Quartile Range (IQR).
- **g**Look at the BMW 5 series. Are there more data in the interval 31-38 or in the interval 45-55? How do you know this?
- **h**Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

  - **i**31-35
  - **ii**38-41
  - **iii**41-64

Exercise:

Problem:

The following box plot shows the U.S. population for 1990, the latest available year. (Source: Bureau of the Census, 1990 Census)



- **a**Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
- **b**12.6% are age 65 and over. Approximately what percent of the population are of working age adults (above age 17 to age 65)?

Solution:

- **a**more children
- **b**62.4%

Exercise:

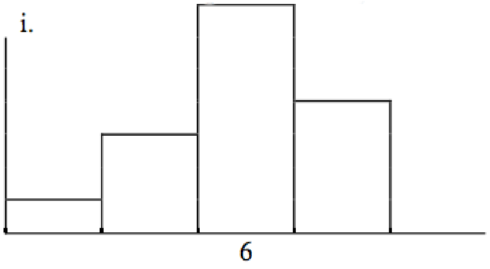
Problem:

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information:

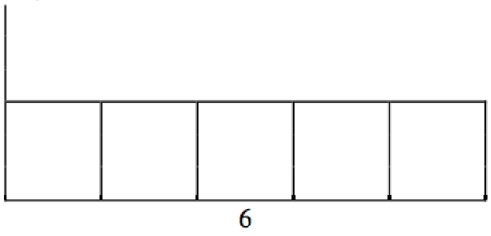
	Javier	**Ercilla**
x	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

- **a**How can you determine which survey was correct ?
- **b**Explain what the difference in the results of the surveys implies about the data.
- **c**If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

i.

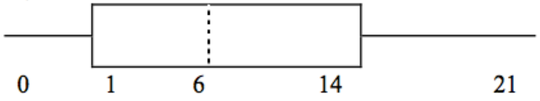


ii.

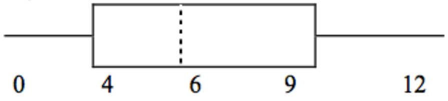


- **d**If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia’s sample? How do you know?

i.



ii.



Exercise:

Problem: Student grades on a chemistry exam were:

77, 78, 76, 81, 86, 51, 79, 82, 84, 99

- **a**Construct a stem-and-leaf plot of the data.
- **b**Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Solution:

- **b**51,99

Try these multiple choice questions (Exercises 24 - 30).

The next three questions refer to the following information. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency

Number of years	**Frequency**
7	1
14	3
15	1
18	1
19	4
20	3
22	1
23	1
26	1
40	2
42	2
	Total = 20

Exercise:

Problem: What is the IQR?

- **A**8
- **B**11
- **C**15
- **D**35

Solution:

A

Exercise:

Problem: What is the mode?

- **A**19
- **B**19.5
- **C**14 and 20
- **D**22.65

Solution:

A

Exercise:

Problem: Is this a sample or the entire population?

- **A**sample
- **B**entire population
- **C**neither

Solution:

B

The next two questions refer to the following table. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Exercise:

Problem: The 80th percentile is:

- **A**5
- **B**80
- **C**3
- **D**4

Solution:

D

Exercise:

Problem:

The number that is 1.5 standard deviations BELOW the mean is approximately:

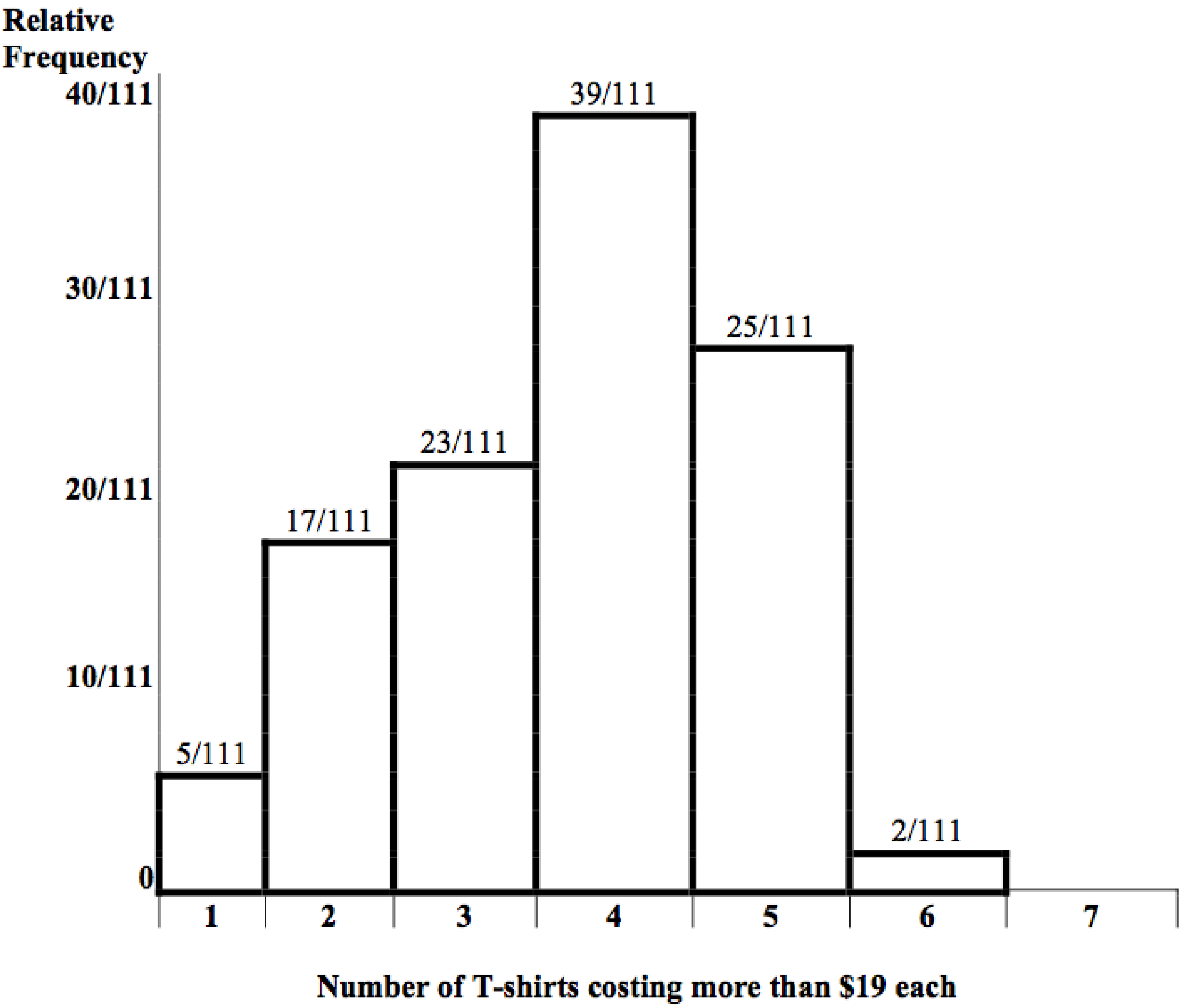
- **A**0.7

- **B**4.8
- **C**-2.8
- **D**Cannot be determined

Solution:

A

The next two questions refer to the following histogram. Suppose one hundred eleven people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



Exercise:

Problem:

The percent of people that own at most three (3) T-shirts costing more than $19 each is approximately:

- **A**21
- **B**59
- **C**41
- **D**Cannot be determined

Solution:

C

Exercise:

Problem:

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

- **A**cluster
- **B**simple random
- **C**stratified
- **D**convenience

Solution:

D

Exercise:

Problem:

Below are the **2010 obesity rates by U.S. states and Washington, DC.**(*Source: http://www.cdc.gov/obesity/data/adult.html)*)

State	**Percent (%)**	**State**	**Percent (%)**
Alabama	32.2	Montana	23.0
Alaska	24.5	Nebraska	26.9
Arizona	24.3	Nevada	22.4
Arkansas	30.1	New Hampshire	25.0
California	24.0	New Jersey	23.8
Colorado	21.0	New Mexico	25.1
Connecticut	22.5	New York	23.9
Delaware	28.0	North Carolina	27.8
Washington, DC	22.2	North Dakota	27.2
Florida	26.6	Ohio	29.2
Georgia	29.6	Oklahoma	30.4
Hawaii	22.7	Oregon	26.8
Idaho	26.5	Pennsylvania	28.6
Illinois	28.2	Rhode Island	25.5
Indiana	29.6	South Carolina	31.5

State	Percent (%)	State	Percent (%)
Iowa	28.4	South Dakota	27.3
Kansas	29.4	Tennessee	30.8
Kentucky	31.3	Texas	31.0
Louisiana	31.0	Utah	22.5
Maine	26.8	Vermont	23.2
Maryland	27.1	Virginia	26.0
Massachusetts	23.0	Washington	25.5
Michigan	30.9	West Virginia	32.5
Minnesota	24.8	Wisconsin	26.3
Mississippi	34.0	Wyoming	25.1
Missouri	30.5		

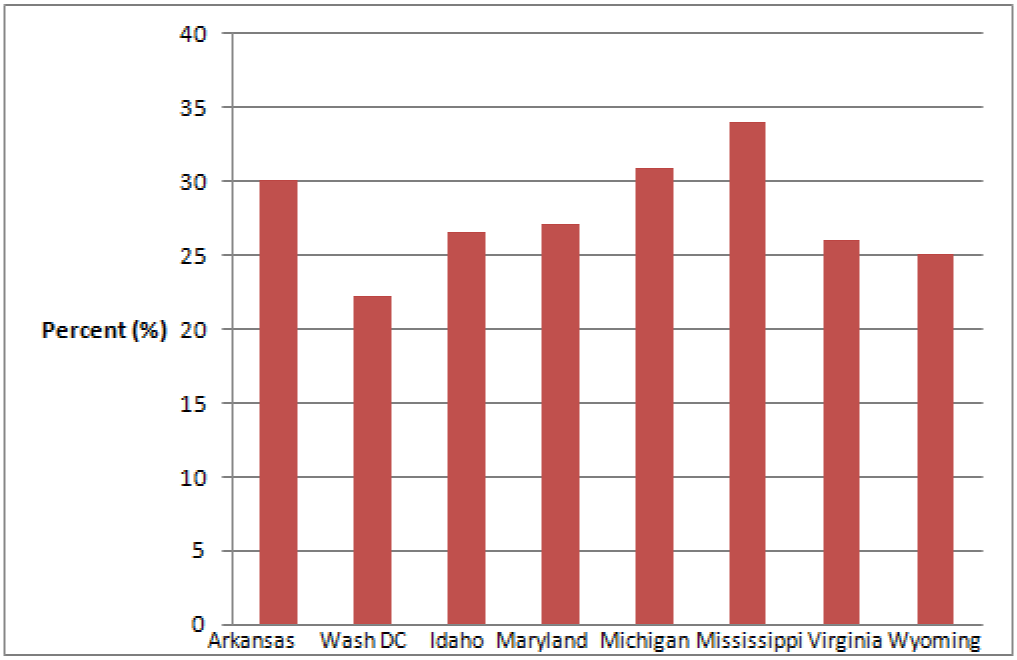
- **a.**Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states.
- **b.**Use a random number generator to randomly pick 8 states. Construct a bar graph of the obesity rates of those 8 states.
- **c.**Construct a bar graph for all the states beginning with the letter "A."
- **d.**Construct a bar graph for all the states beginning with the letter "M."

Solution:

Example solution for **b** using the random number generator for the Ti-84 Plus to generate a simple random sample of 8 states. Instructions are below.

- Number the entries in the table 1 - 51 (Includes Washington, DC; Numbered vertically)
- Press MATH
- Arrow over to PRB
- Press 5:randInt(
- Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}. Corresponding percents are {28.7 21.8 24.5 26 28.9 32.8 25 24.6}.



Exercise:

Problem:

A music school has budgeted to purchase 3 musical instruments. They plan to purchase a piano costing \$3000, a guitar costing \$550, and a drum set costing \$600. The mean cost for a piano is \$4,000 with a standard deviation of \$2,500. The mean cost for a guitar is \$500 with a standard deviation of \$200. The mean cost for drums is \$700 with a standard deviation of \$100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer numerically.

Solution:

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar cost the most in comparison to the cost of other instruments of the same type.

Exercise:

Problem:

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the table below. (Note that this is the data presented for publisher B in homework exercise 13).

# of books	Freq.	Rel. Freq.

# of books	**Freq.**	**Rel. Freq.**
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher B

a. Are there any outliers in the data? Use an appropriate numerical test involving the IQR to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than 2 standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts (a) and (c) of this problem give the same answer?
e. Examine the shape of the data. Which part, (a) or (c), of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Solution:

- IQR = 4 – 1 = 3 ; Q1 – 1.5*IQR = 1 – 1.5(3) = -3.5 ; Q3 + 1.5*IQR = 4 + 1.5(3) = 8.5 ;The data value of 9 is larger than 8.5. The purchase of 9 books in one month is an outlier.
- The outlier should be investigated to see if there is an error or some other problem in the data; then a decision whether to include or exclude it should be made based on the particular situation. If it was a correct value then the data value should remain in the data set. If there is a problem with this data value, then it should be corrected or removed from the data. For example: If the data was recorded incorrectly (perhaps a 9 was miscoded and the correct value was 6) then the data should be corrected. If it was an error but the correct value is not known it should be removed from the data set.
- xbar – 2s = 2.45 – 2*1.88 = -1.31 ; xbar + 2s = 2.45 + 2*1.88 = 6.21 ; Using this method, the five data values of 7 books purchased and the one data value of 9 books purchased would be considered unusual.
- No: part (a) identifies only the value of 9 to be an outlier but part (c) identifies both 7 and 9.
- The data is skewed (to the right). It would be more appropriate to use the method involving the IQR in part (a), identifying only the one value of 9 books purchased as an outlier. Note that part (c) remarks that identifying unusual data values by using the criteria of being further than 2 standard deviations away from the mean is most appropriate when the data are mound-shaped and symmetric.
- The data are skewed to the right. For skewed data it is more appropriate to use the median as a measure of center.

**Exercises 32 and 33 contributed by Roberta Bloom